

# Gegenees V 1.1.0

What is Gegenees?.....	1
Version system:.....	2
What's new.....	2
Installation:.....	2
Perspectives.....	4
The workspace.....	4
The local database.....	6
Populate the local database.....	7
Gegenees genome format.....	7
Gegenees comparisons.....	8
Creating a fragmented all-all comparison.....	9
The alignment.....	10
The analysis .....	10
Included genomes tab.....	11
Group settings tab.....	11
Heat plot tab.....	12
Score overview tab.....	13
Viewing a signature in Artemis.....	15
Score table tab.....	16
Primer mapping/Primer score table tab.....	22

## What is Gegenees?

Gegenees is a software that compares genome sequences (Draft and Completed). It was primarily developed for bacterial genomes but it is also possible to use on viruses and smaller eukaryotes. Gegenees fragments the genomes and compares all pieces against all genomes. Based on this all-against-all comparison, a phylogenetic data can be extracted. It is also possible to define a "target group" and search for genomic regions that have high specificity for the target group. This is referred

to as a "genomic signature". The genomic signature regions can be used to find candidate regions for the design of primers and probes for new highly specific diagnostic assays. There is also a built in primer/probe verification system that compares new candidate or existing primers and probes to the genomic database and to the defined target groups. Future versions will include more aspects of comparing next generation sequencing (NGS) data.

## Version system:

The first released version was 1.0.1. Based on user feedback, new versions will be released that solves problems and makes the program easier to use. These versions will be called 1.0.1, 1.0.2 ....

New functionality will lead to version 1.1.1, 1.2.1 ....

To see your version, select "About Gegenees" in the Help main menu.



## What's new

Since version 1.0.4 our aim has been to make Gegenees more intuitive and user friendly. The implementation of wizards for creating comparisons was one major step in this direction, but also the new database manager and ftp client along with some "under the hood" structural rearrangements facilitates this aim.

## Installation:

The software can be downloaded from <http://www.gegenees.org>. There are several variants uploaded, depending on your operating system (OS). Windows. Macintosh and Linux are supported. You must also chose the correct processor architecture (32 bit or 64 bit).

To see the architecture :

In Windows 7, select "properties" when right clicking "computer".

In Macintosh, select " About this Mac" from the Apple menu. Mac OS X 10.5 (or greater) is a 64-bit.

In Linux, in a terminal, type "uname -a". If "x86\_64" or "ia64" is shown, the system is 64 bit.

Java needs to be installed. You may check your Java version at this link:

<http://java.com/en/download/installed.jsp>.

Download and extract the compressed Gegenees program. Run the Gegenees program from in the "Gegenees" folder. You will then be asked to specify a "Workspace". This is the place where all your genomes and comparisons will be stored.

If Gegenees starts OK, it is time to install BLAST+. Gegenees can download and install BLAST+ for you if you select "Software" and then click on Download BLAST+. BLAST+ is then downloaded and installed in your Gegenees installation folder under the new directory /bin/blast/.

BLAST+ can also be downloaded from <ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/>

In general, the latest version of BLAST+ is recommended and can be downloaded from

<ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST>

There is one exception though, and it is regarding Windows computers. At the time of writing, the latest version of BLAST+ is version 2.2.26, which for unknown reasons and seemingly at random gets stuck. Version 2.2.25 works fine though and this is the version that Gegenees uses in its automatic download.

Chose a version that corresponds to your OS and architecture (32 bit/64 bit) and extract (or in windows run the installation program).

If you have downloaded BLAST+ manually, Gegenees must be configured to find BLAST. Click "File" and then "Preferences" in the Gegenees menu. In the preference dialog click on "Third-party software components" and specify the full path name to the directory containing the BLAST+ executables and then click "OK". E.g. Windows, C:\blast+\bin\ or Linux, \usr\local \blast\

If BLAST+ is added to your system path, you may not need to specify the path.



A function for testing if Gegenees finds blast is under implementation and will come in the next version.

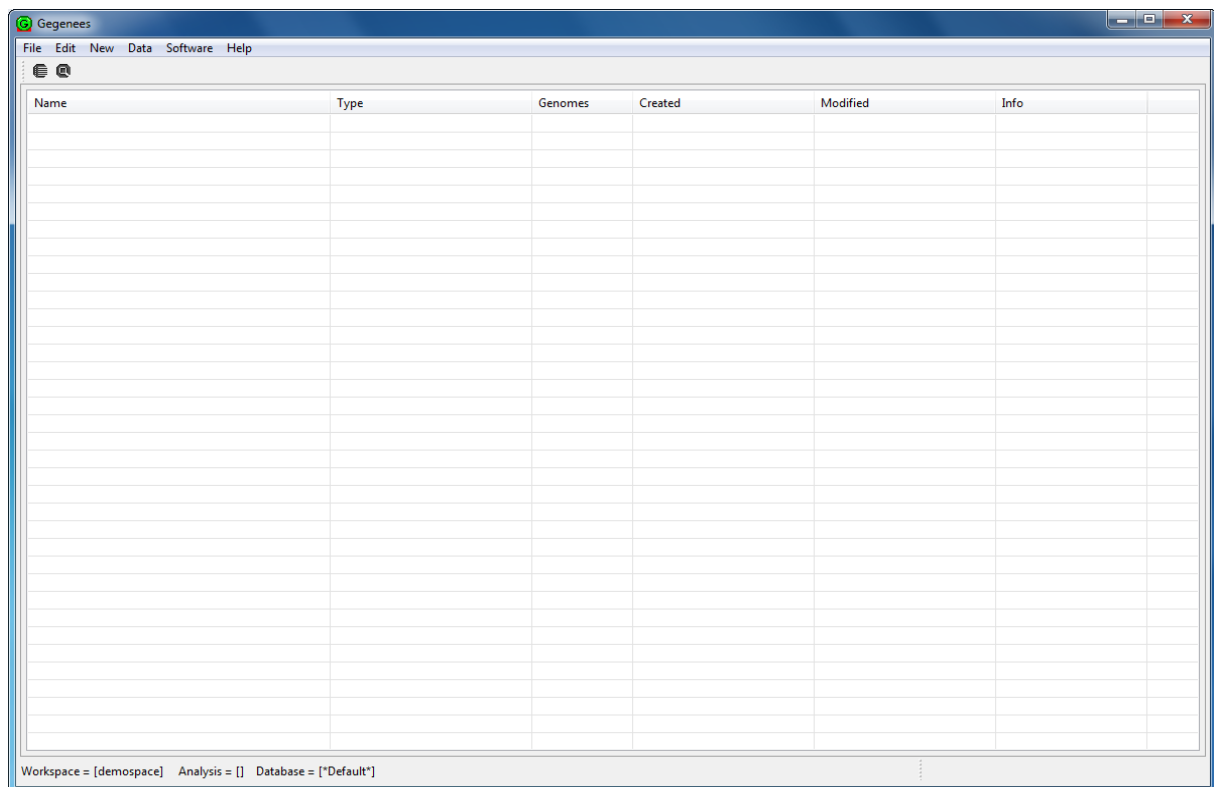
## Perspectives

Gegenees have two main "perspectives". The active perspective can be changed in the "perspective bar" just under the main menu. The perspectives are:

- Workbench overview:** An overview of all comparisons collected in this "workspace".
- Analysis:** a perspective where the comparative calculations are started and controlled but also for phylogenomic and signature analysis of a completed alignment.

## The workspace

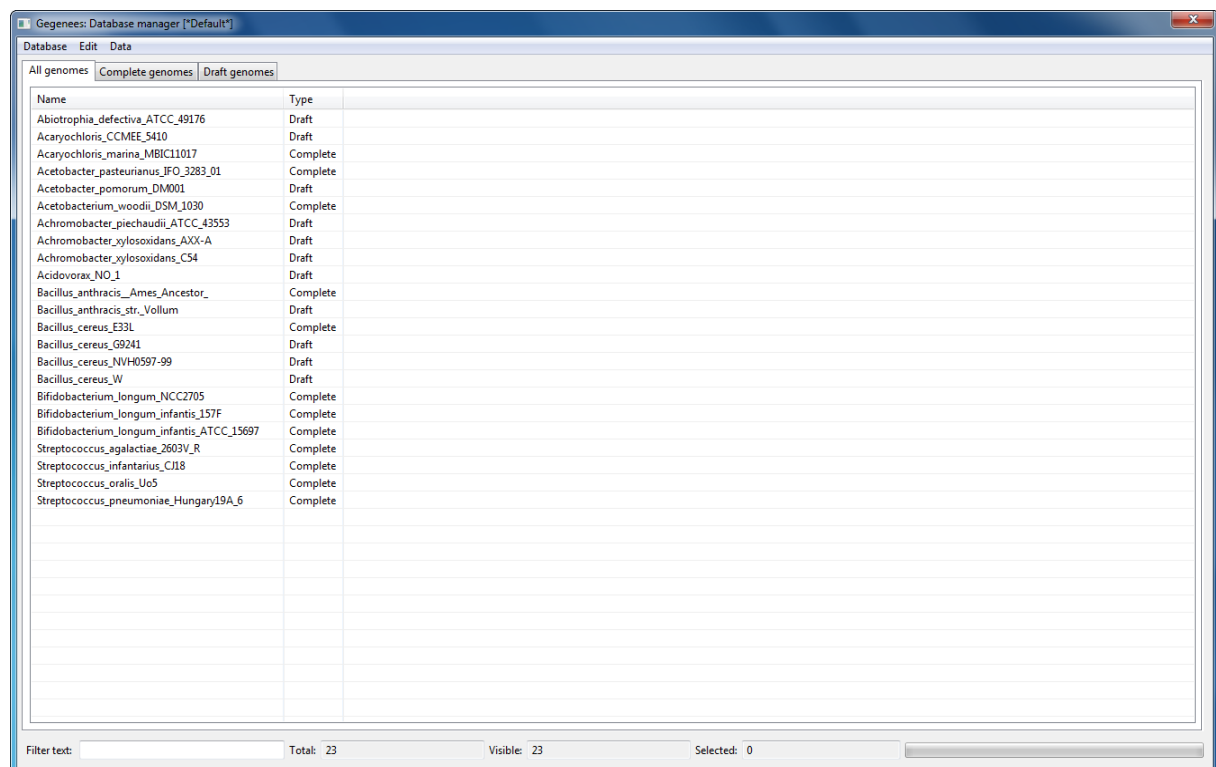
The workspace is where all your genomes and your comparison projects will be stored. A workspace must always be selected so when Gegenees starts for the first time, a workspace location is asked for. Typically, a user always use the same workspace. Different users on the same computer might want to collect their genomes and comparisons in their own, separate workspace. The workspace can be changed by clicking "File" and "Switch workspace". The name of the current workspace can always be seen at the bottom left side of the "status line". The comparisons that are present in the current workspace will be listed in the table in the Workbench overview perspective.



## The local database

This is where you store your genomes. A "default" database is always present but customized databases can also be made. The default database represents a directory called "database" in the workspace. Custom databases represent directories named "database\_NameOfCustomDatabase".

The local database content is shown by clicking “Data” and then “Database manager...”. If the database is large, subsets can be shown by entering a case sensitive filter text. (e.g Bacillus to show only the genus Bacillus). There is also a filter for showing only genomes in draft or complete form.



The screenshot shows the 'Gegenees: Database manager ["Default"]' window. It has a menu bar with 'Database', 'Edit', and 'Data'. Below the menu bar are three tabs: 'All genomes', 'Complete genomes', and 'Draft genomes'. The 'All genomes' tab is selected, displaying a table with two columns: 'Name' and 'Type'. The table lists various bacterial genomes, including Abiotrophia\_defectiva, Acaryochloris, Acetobacter, Acetobacterium, Achromobacter, Acidovorax, Bacillus, Bifidobacterium, and Streptococcus. The 'Type' column indicates whether each genome is 'Draft' or 'Complete'. At the bottom of the window, there is a 'Filter text:' field, and status information showing 'Total: 23', 'Visible: 23', and 'Selected: 0'.

Name	Type
Abiotrophia_defectiva_ATCC_49176	Draft
Acaryochloris_CCMEE_5410	Draft
Acaryochloris_marina_MBI11017	Complete
Acetobacter_pasteurianus_IFO_3283_01	Complete
Acetobacter_pomorum_DM001	Draft
Acetobacterium_woodii_DSM_1030	Complete
Achromobacter_piechaudii_ATCC_43553	Draft
Achromobacter_xylosoxidans_AXA-A	Draft
Achromobacter_xylosoxidans_C54	Draft
Acidovorax_NO_1	Draft
Bacillus_anthraxis_Ames_Ancestor_	Complete
Bacillus_anthraxis_str_Vollum	Draft
Bacillus_cereus_E33L	Complete
Bacillus_cereus_G9241	Draft
Bacillus_cereus_NVH0597-99	Draft
Bacillus_cereus_W	Draft
Bifidobacterium_longum_NCC2705	Complete
Bifidobacterium_longum_infantis_157F	Complete
Bifidobacterium_longum_infantis_ATCC_15697	Complete
Streptococcus_agalactiae_2603V_R	Complete
Streptococcus_infantarius_CJ18	Complete
Streptococcus_oralis_Uo5	Complete
Streptococcus_pneumoniae_Hungary19A_6	Complete

## Populate the local database

It is possible to download genomes from a remote FTP server by clicking in the database manager “Data” and then “Import by FTP...”. A few predefined FTP sites are distributed with the Gegenees software and can be found in the section called “FTP client settings” at the center of the window. These predefined sites are the “NCBI complete genomes” (pointing at <ftp://ftp.ncbi.nih.gov/genomes/Bacteria/>) and “NCBI genomes bacteria draft” (pointing at [ftp://ftp.ncbi.nih.gov/genomes/Bacteria\\_DRAFT/](ftp://ftp.ncbi.nih.gov/genomes/Bacteria_DRAFT/)) and “NCBI Genbank bacteria draft” (pointing at <ftp://ftp.ncbi.nih.gov/genbank/wgs/>). “NCBI Genbank bacteria draft” contains more genomes than “NCBI genomes bacteria draft”. Select one in the combo box and click on the button with two yellow arrows to connect. When you are connected to the FTP server you get a list of available genomes where you can select the ones you want to download. When you have selected the genomes you wish to download click the green arrow to download them to your local database.

The FTP-sites are defined by a file ending with “.site” in the “ftp” directory in the workspace directory. It is possible to make own “.site” files. Copy the content of the existing files and replace the “HostName:” and the “Directory:” fields with custom information. The FTP site must be formatted as the NCBI site. If you need to connect to other types of ftp sites, contact the support at Gegenees.

## Gegenees genome format

Gegenees uses a folder for each genome. The folder name corresponds to the Name of the genome as appearing in Gegenees and the type of the genome (“Draft”/“Complete”). The last part of the folder name has the format “--Draft--” or “--Complete--”. In future versions, more types may be

introduced. The folder contains at least one Genbank formatted file with extension ".gbk" or ".gbff". If there are several subsequences/contigs, they may be in the same file or in separate files. Genomes are stored in the database folder(s) but also in the comparison folders. Thus, the comparison keeps a copy of the part of the database it is using. If the database is modified, the genomes belonging to a comparison is still untouched in the comparison folder.

## Gegenees comparisons

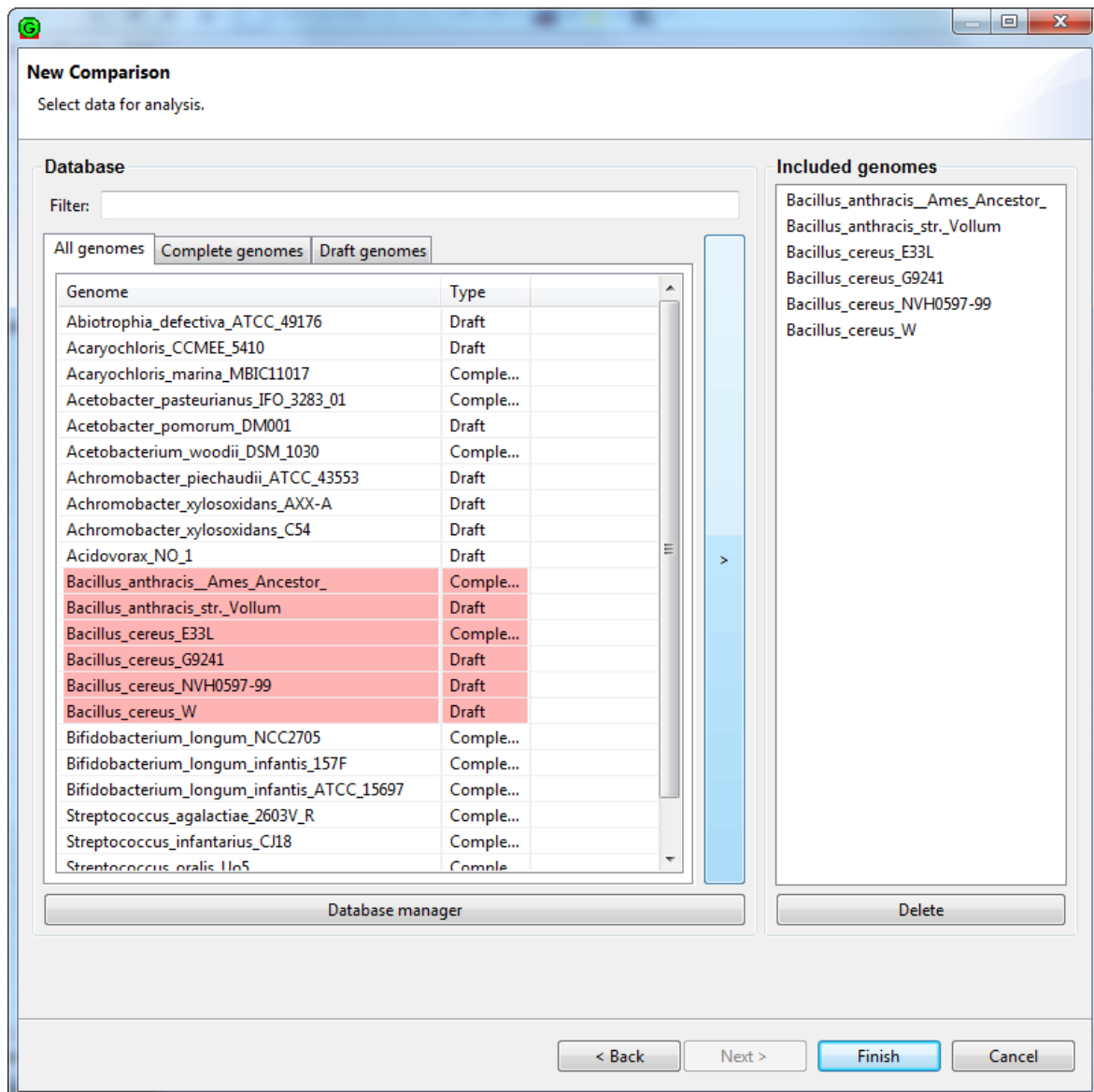
At present, Gegenees conducts two base types of analyses; a “fragmented all against all comparison” and a “primer mapping”. Structurally, Gegenees treats all analyses as comparisons even if it is called primer mapping. The Gegenees "comparison" is a "package" containing:

- a list of genomes that belongs to the comparison
- a copy of all the genomes as they looked like when they were added to the comparison
- one "alignment" which represents a comparative calculation at a certain resolution (see below).
- Files coupled to the analysis of the alignment.

A comparison is represented in the computer's file system as a folder with the prefix "comparison\_" and it contains copies of the genomes and a file "comparison.geg" that contains the genome list.

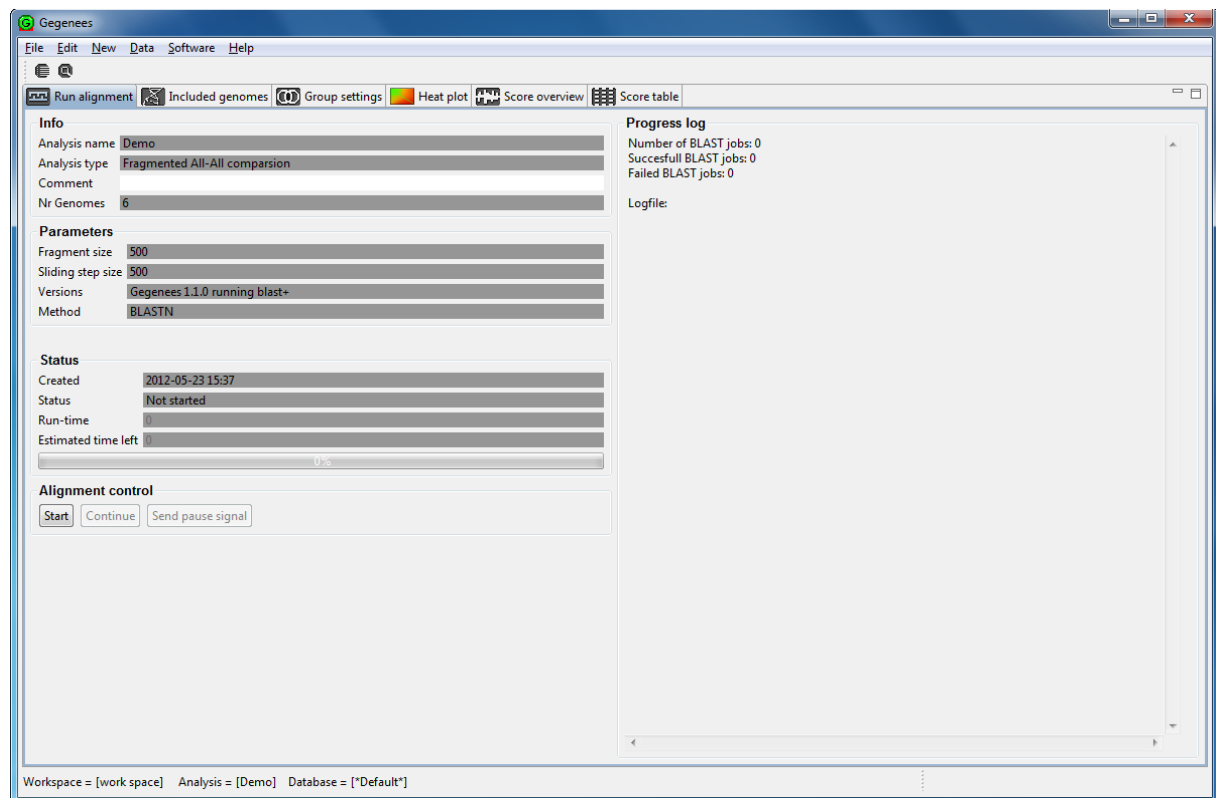
## Creating a fragmented all-all comparison

To create a fragmented all-all comparison, click on “New” and then “Fragmented all-all comparison” in the main menu. A wizard will help you set up the comparison by first letting you chose a name and some alignment settings. The resolution of the alignment is controlled by two parameters, the "Fragment size" and the "Sliding step size". The fragment size represents the "scanning window size" and it should be smaller than the genomic region you anticipate to find in the analysis. For bacteria, we recommend 200/100 (frag-size /slide-size ) which is more accurate and 500/500 (much faster and usually sufficient) settings. Small fragment sizes and sliding step sizes gives more demanding calculations. When working with viruses and small sequences, shorter settings may be needed. It is also possible to use tblastx (compares sequences on translated level, i.e. amino acids). This is much more demanding and the datasets should be smaller. If the sequences are pylogenetically far apart, this may be a useful operating mode. Then you have to select genomes from your database to include in the comparison and click “Finish”. By clicking finish you will be taken to the analysis perspective.



## The alignment

When you which to start the alignment process, click the start button. The calculation progress will be shown and the log-window in the right part will show messages on what is happening. Typically, first a lot of conversion and preparation messages appears. Then the a BLAST list is created and executed in parallel "threads". Typically, each "thread" should not take more than at the most a few minutes to complete. The number of simultaneously calculating threads is indicated and also the thread number and the total number of threads that should be run. It is possible to send a pause signal and then resume the calculation later. After all the threads have been run, some data analysis is made and then the alignment is completed. Once an alignment is completed, it is possible to analyze the data in the other tabs in the analysis perspective. An alignment is represented on the hard drive by a folder with the prefix "alignment\_analysis".



## The analysis

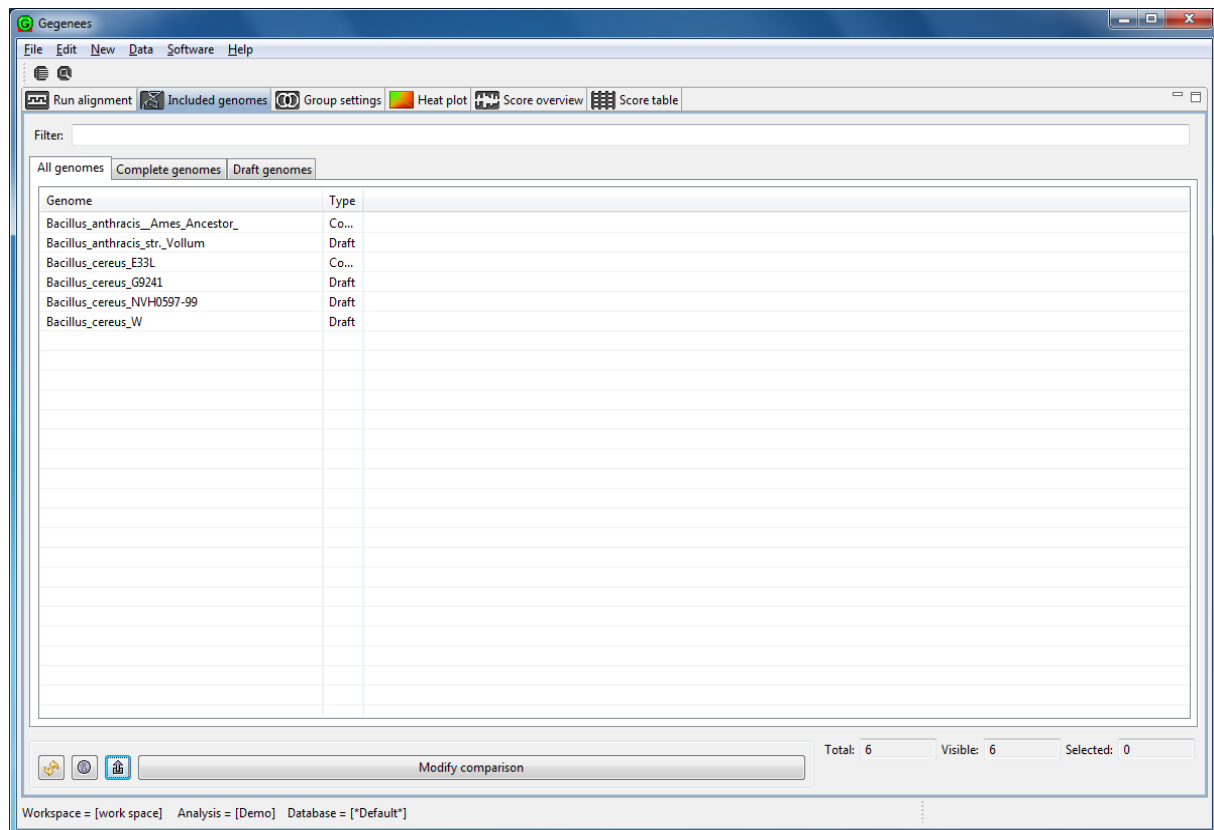
The analysis perspective also contains different tools for analyzing the data from a completed alignment. The tabs are:

- Included genomes: This tab is a list of all genomes included in the analysis. There is also the place for add or remove genomes from the comparison.
- Group settings: A tabular view of the genomes that allows definition of the target and background groups for the analysis.
- Heat plot: This tab represents an phylogenomic overview of the data based on average similarities.
- Score overview: This tab represents a graphical overview and exploration tool of the genomic regions that are unique or discriminatory for the target group.
- Score table: This tab represents a tabular exploration tool for the genomic regions that are unique or discriminatory for the target group.

## Included genomes tab

The included genomes tab shows all the genomes that are included in the comparison. If you want to add or remove genomes from your comparison it is possible by clicking on the button “Modify comparison”. It is also possible to export genome info for all genomes that are included in the comparison.



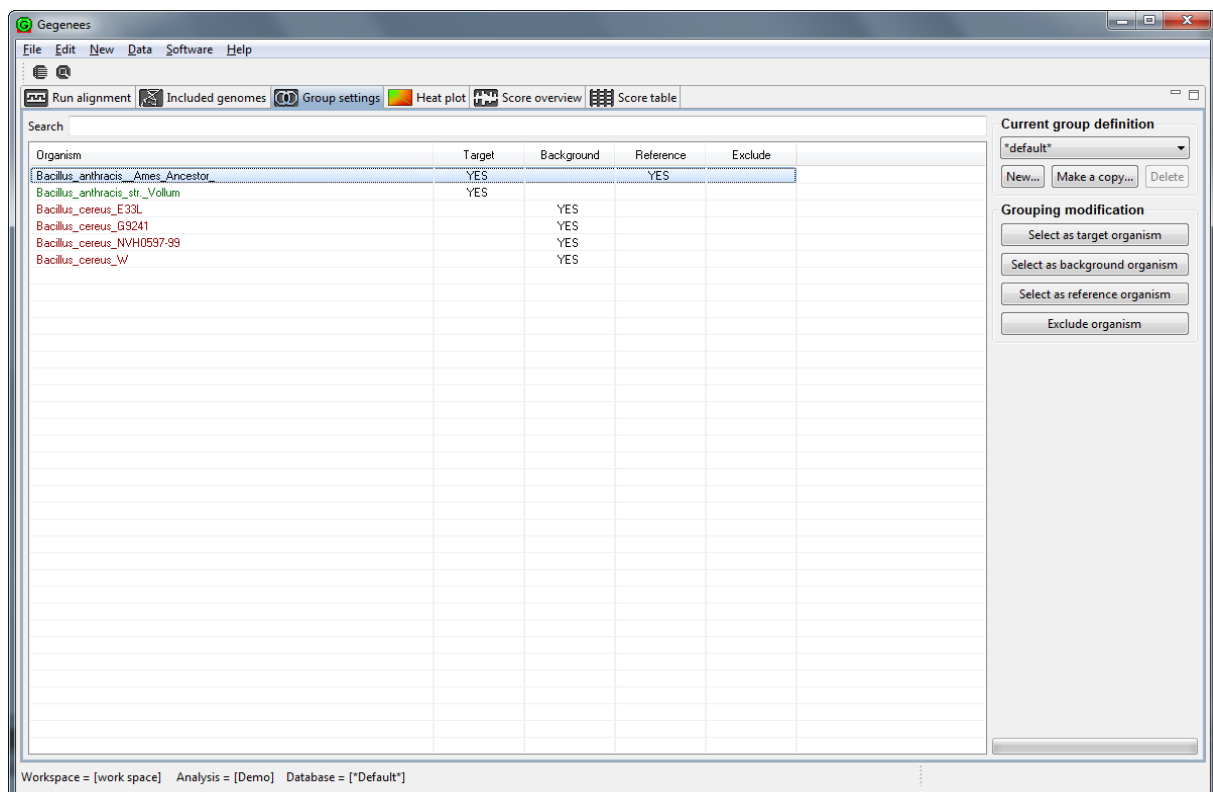


## Group settings tab

The group settings tab allows the target group and the background group to be defined.

- The target group represents the genomes that you are interested in. You might want to find discriminatory genomic regions that can be used to design specific molecular assays. You might also be interested in finding and analyzing genomic regions that are associated with a specific phenotypic trait that the target group share. The target group should ideally contain genomes that represents the genomic variability in the group as well as possible.
- The background group represents other genomes. The background should contain the genomes that are most likely to give a cross-reaction (false positive in an assay). For best result, close neighbor strains/isolates that do not contain the phenotypic trait should be included.
- Excluded genomes. You may also chose to exclude genomes from the analysis.
- The reference genome is the genome where the nucleotide coordinate system and the annotations are collected from. For a high stringency biomarker analysis (see below), the best annotated genome in the target group should be used. For lower stringency in the biomarker analysis, the result may be slightly different depending on which target genome is selected as reference. It is then good to look at the data with different reference genomes. **A reference genome must be chosen to be able to look at the score overview and score table tabs.**

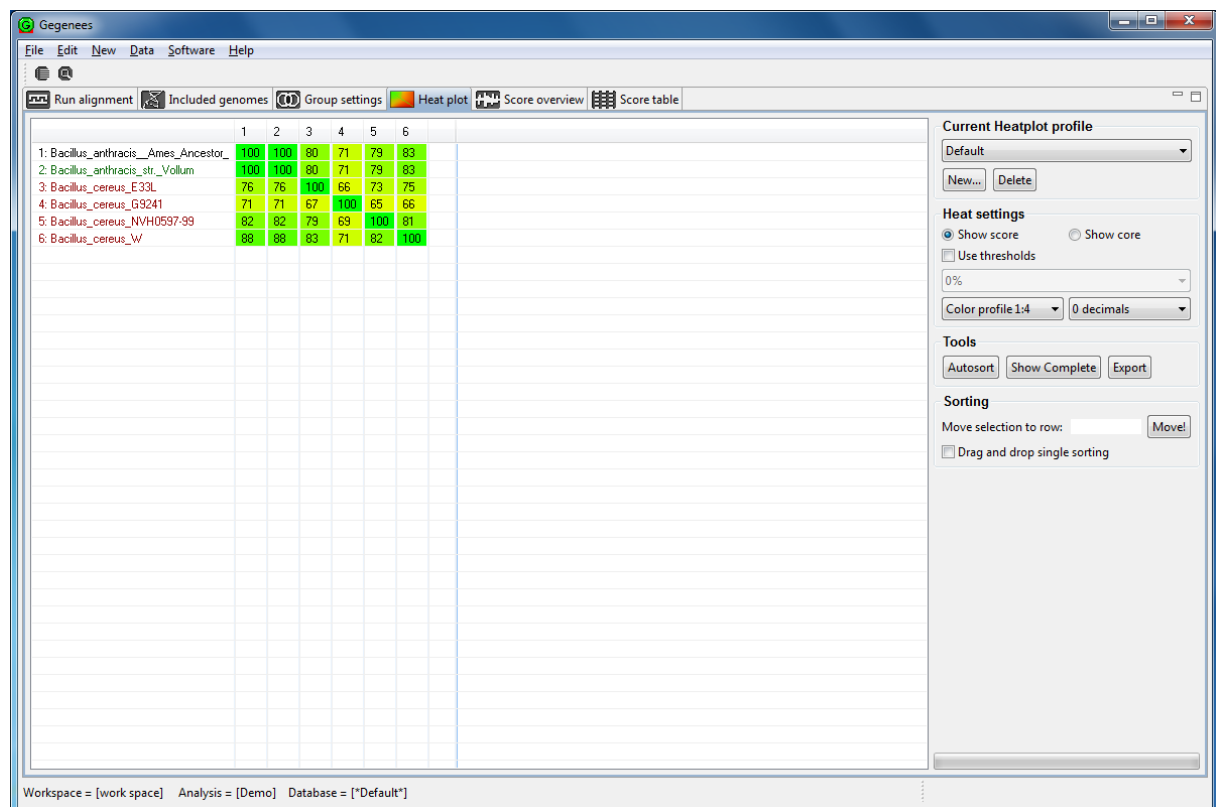
Several group settings can be created from the same dataset (e.g. different subtypes) with the "New.." or "Make a copy.." buttons.



## Heat plot tab

The heat plot tab gives an phylogenomic overview of the data. It is the average normalized BLAST score values of all fragments that are shown. It is also possible to define threshold values, meaning that that fragments falling under the threshold is not used to calculate the average similarity value. This gives a better phylogenetic signal since the similarity value is only based on conserved genetic material (the core genome). It is also possible to see how large the core genome is at the specified threshold (select "show core" instead of "show score"). It is possible to change the "color profile" of the heat plot so that differences are highlighted as well as possible for the particular dataset. The number of decimals shown can be changed. The genomes are sorted alphabetically, which often is sufficient. There are sorting possibilities built in for the heat plots. It is possible to move genomes or group of genomes with the "Move selection to row" field or by right clicking it and select "move" from the context menu. The target and background group settings can also be modified from the right click-context menu. If, "Drag and drop, single sorting" is selected, genomes can be dragged with the mouse, one by one. The sort is saved and if several sorts are wanted, new ones can be created with the "new..." button. There is also an "autosort" function, that tries to minimize the score distances between the rows. There is also an export button that allows of the phylogenomic data:

- export of the table in tab-format (for work in spreadsheet programs)
- export as html. Can be opened in a web browser or converted to publication grade figures. This is sometimes a better overview if the table is very large.
- export nexus file. Use this export to create dendrograms in e.g. [SplitsTree](#) .



## Score overview tab

The score overview tab shows a graphic representation of the "biomarker scores". Biomarker scores are score values that rank all genomic regions (fragments) in how discriminating they are for the target group in terms of conservation (no false negatives) and uniqueness (no false positives in the background). There are three types of scores with different stringency:

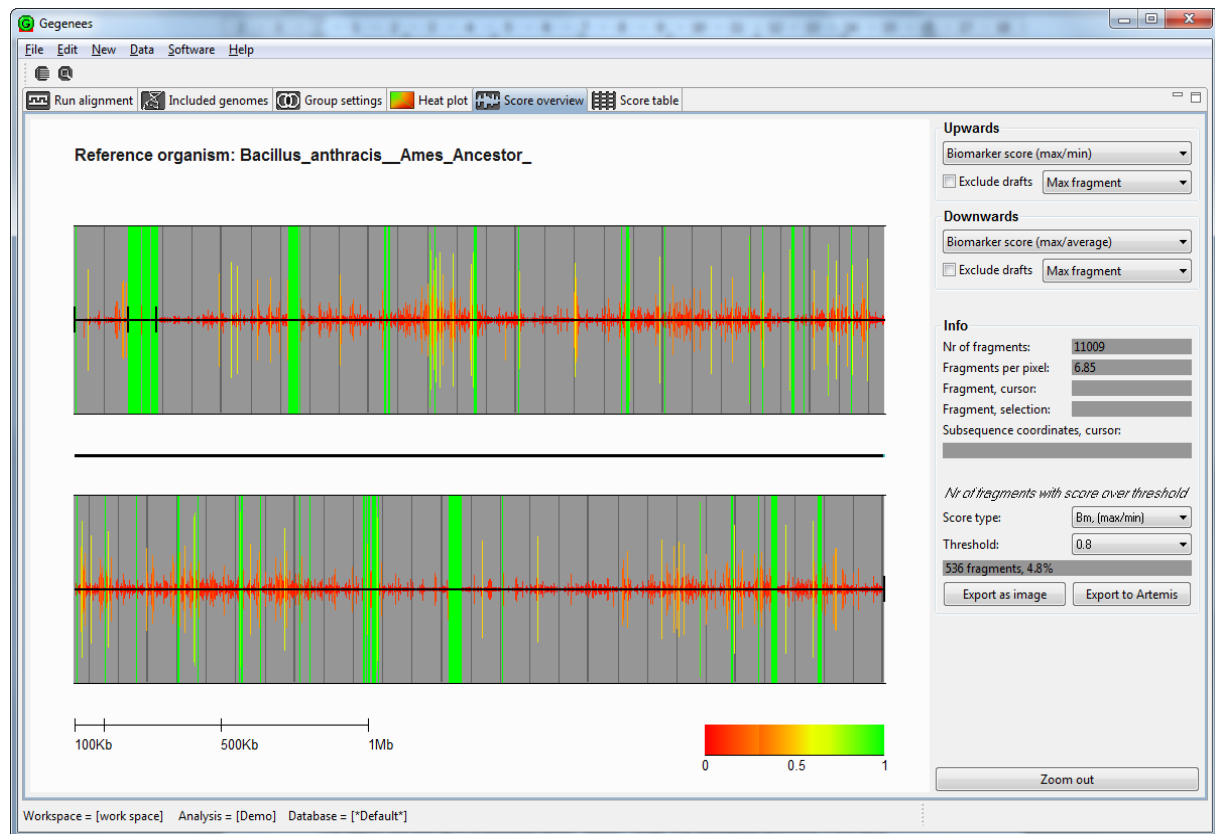
- |                                   |   |
|-----------------------------------|---|
| Biomarker score (max/min):        | represents the highest stringency. The high scoring fragments must be present and conserved in ALL target group genomes and must be absent or very diverged in ALL background genomes. The score goes from 0 (or negative values) representing bad regions up to 1 which represents a perfect region. The score is based on the worst background genome (max score) and on the worst target group genome (min score). |
| Biomarker score (max/average)     | Similar to Biomarker score (max/average) but uses average values of the background group (relaxes the uniqueness criterion)   |
| Biomarker score (average/average) | Similar to Biomarker score (max/average) but uses average values of the background group and the target group (relaxes the uniqueness and the conservation criterion)   |
| Target group maximum score        | Shows the best score in the target group (self score). This is usually 100% everywhere, but if bad regions are present in the sequence (e.g. nnnnnnnnnnnnnn), they can be identified here.  |

Target group minimum score	Shows the worst conservation within the target group. Can be used to find highly conserved regions.
Target group average score	Shows the average conservation within the target group. Can be used to find highly conserved regions.
Background group maximum score	Shows the best score in the background group. This represent the worst cross reaction.
Background group minimum score	Shows the minimum score in the background group. Can be used to find regions that are not conserved in a group of very related genomes.
Background group average score	Shows the average conservation within the background group.

The biomarker scores are drawn graphically. There is possibilities to compare two types of scores by drawing one upwards (from the coordinate axis) and a second downwards. The graphical view is spitted into two rows in order to use the computer screen optimal (do not confuse the upper row with the upwards score, there are upwards scores in both rows) . there is a possibility to exclude draft genomes in the calculations since they sometimes lack regions that in some cases may disturb the analysis. When the mouse moves over the graph, the sub-sequence, fragment number and coordinate at the cursor position is shown in the "info" part to the right. The number of fragments that each pixel column on the screen represent, is also indicated. It is also possible to see how many percent of the genome has a biomarker score over a certain threshold. This gives a good overview of the how much genomic regions one can expect.

It is possible to zoom into the graph by selecting a region with the right mouse button down. If the left mouse button is used, the corresponding region is "selected". A selection can thereafter be loaded into the tabular view for further data mining by a right click.

There are possibilities to export the graph (as seen on the screen) as an image. It is also possible to export the data to a file that can be explored in Artemis (see section below).



## Viewing a signature in Artemis

It is possible to export an interesting sub-sequence from the genome (or the whole genome if it is completed) into a format that can be viewed in Artemis

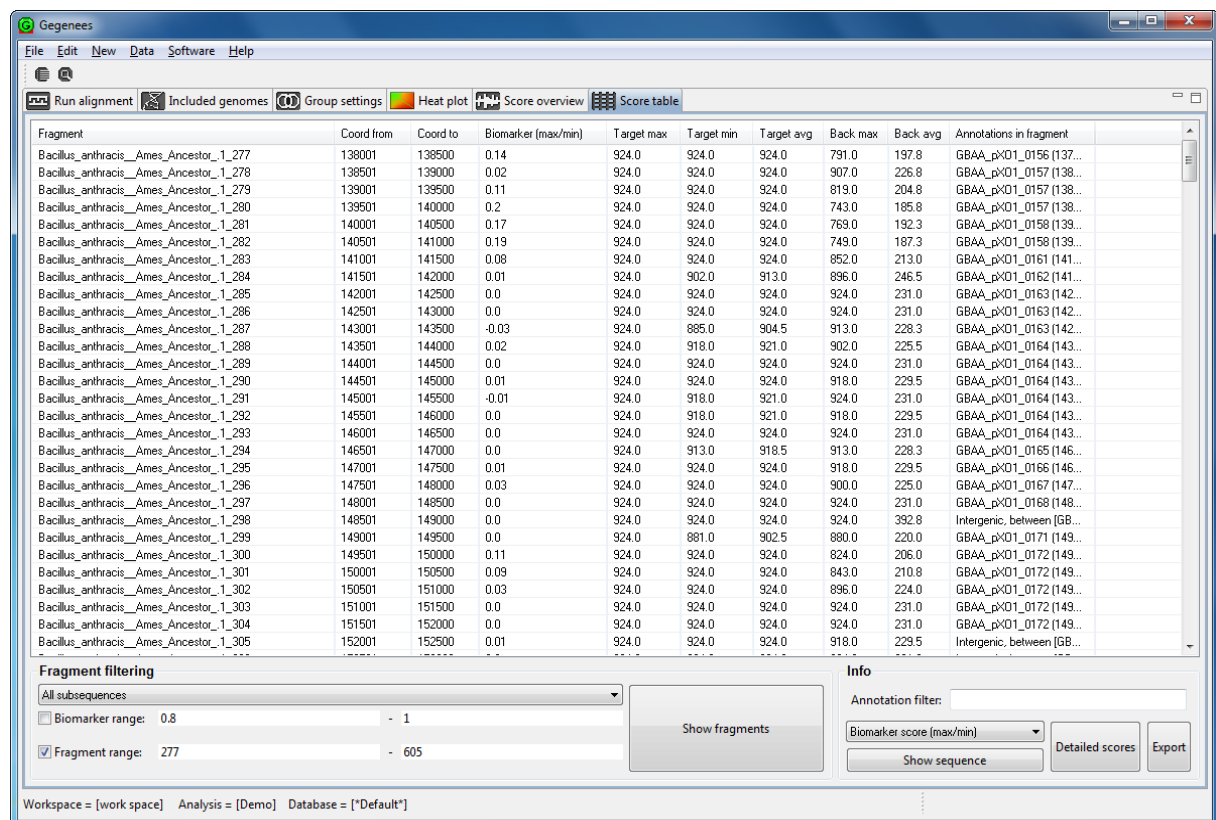
<http://www.sanger.ac.uk/resources/software/artemis/>.

The export will end up in a directory called "export" under the workspace directory. It will be a "\*.gbk" file that essentially is the same file as the original "gbk" file (if there are problems or warnings when loading the original file in Artemis they will remain). The "gene" and "misc feature" track is replaced by the biomarker scores. Five files are exported

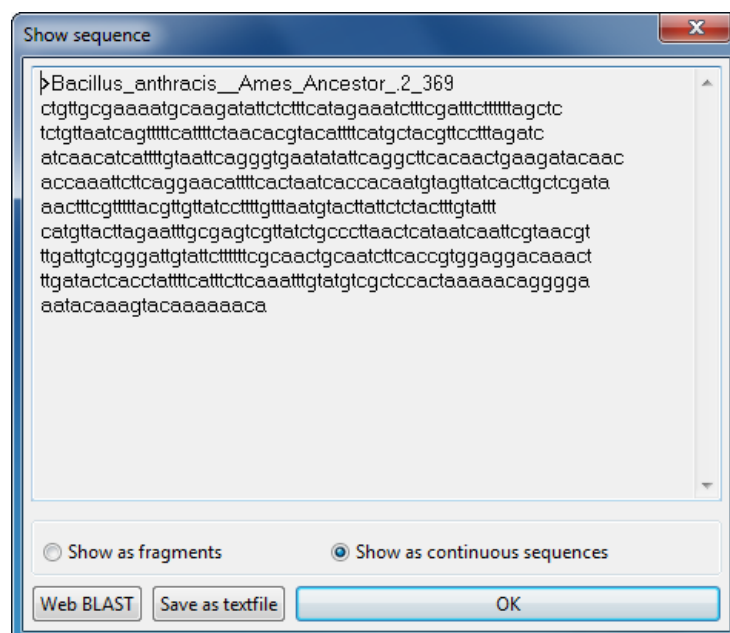
1. the original annotated file with also the gene and misc feature track intact.
2. a file with Biomarker scores (max/min) as a misc\_feature track
3. a file with Biomarker scores (max/avr) as a misc\_feature track
4. a file with Biomarker scores (avr/avr) as a misc\_feature track
5. a file with Biomarker scores (max/min), only complete genomes as a misc\_feature track

An example of how this kind of export will look like in Artemis is shown below.





Show sequence displays the actual sequences of the fragments and it is possible to fuse adjacent and overlapping fragments into continuous sequences. The sequences can be exported to a Fasta-file or sent to a web page ready for a blast comparison at NCBI.



The detailed scores, shows how each fragment scores against each genome in the target and background group. This may help to identify which particular strain is causing a cross reaction.

Genome	Bacillus_anthraxis_Ames.1_3451	Bacillus_anthraxis_Ames.1_3452	Bacillus_anthraxis_Ames.1_3453	Bacillus_anthraxis_Ames.1_3454	Bacillus_anthraxis_Ames.1_3455	Bacillus
Bacillus_cereus_01	0.0	0.0	0.0	0.0	0.0	0.0
Bacillus_cereus_E33L	0.0	0.0	0.0	0.0	0.0	0.0
Bacillus_cereus_AH820	0.0	0.0	0.0	0.0	0.0	0.0
Bacillus_anthraxis_Sterne	370.0	370.0	370.0	370.0	370.0	370.0
Bacillus_anthraxis_Ames	370.0	370.0	370.0	370.0	370.0	370.0
Bacillus_anthraxis_A0193	370.0	370.0	370.0	370.0	370.0	370.0

It is also possible to export the table (as its shown) or the full data table (without filtering) as a tab delimited text file for further analysis in e.g. a spreadsheet program.

## Primer mapping/Primer score table tab

To make a primer/probe mapping, click on “New” in the top menu and then “Primer mapping...”. Chose a name and enter your primers, click next and select what genomes to run your primer mapping against and run the analysis.

**New primer mapping**  
Chose a name and define your primer sequences for the primer mapping.

Chose a name: DemoPrim

```

>test primer 1
agcgctgca

>test primer 2
ttgatctaagtgaaggat

>test primer 3
acgcgctaacga
  
```

Import primer sequence from FASTA file

< Back Next > Finish Cancel

The primer mapping can now be explored in the "Primer score table tab". The "unalignment" index, represents mismatches in the alignment plus non-aligned nucleotides. It is marked green if it is a perfect match. A target group/ background group setting can be loaded from a fragmented



alignment, and the genomes are color coded accordingly, so that a the primer matching can easily be compared to the target group definition.

Workspace = [work space] Analysis = [DemoPrimer] Database = ["Default"]

Database	Primer name	Unalignment index	Identity	Query length	Alignment length	Mismatches	Gaps
Bacillus_cereus_W--Background--	test_primer_1	3	100.0%	15	12	0	0
Bacillus_cereus_W--Background--	test_primer_2	2	90.0%	20	20	2	0
Bacillus_cereus_NVH0597.99--Background--	test_primer_1	3	100.0%	15	12	0	0
Bacillus_cereus_NVH0597.99--Background--	test_primer_2	0	100.0%	20	20	0	0
Bacillus_cereus_G9241--Background--	test_primer_1	3	100.0%	15	12	0	0
Bacillus_cereus_G9241--Background--	test_primer_2	0	100.0%	20	20	0	0
Bacillus_cereus_E33L--Background--	test_primer_1	3	100.0%	15	12	0	0
Bacillus_cereus_E33L--Background--	test_primer_2	0	100.0%	20	20	0	0
Bacillus_anthraxis_str_Vollum--Target--	test_primer_1	1	93.33%	15	15	1	0
Bacillus_anthraxis_str_Vollum--Target--	test_primer_2	2	90.0%	20	20	2	0
Bacillus_anthraxis_Ames_Ancestral--Reference--	test_primer_1	1	93.33%	15	15	1	0
Bacillus_anthraxis_Ames_Ancestral--Reference--	test_primer_2	2	90.0%	20	20	2	0
Acidovorax_NO_1--Not in group--	test_primer_1	4	100.0%	15	11	0	0
Acidovorax_NO_1--Not in group--	test_primer_2	7	93.33%	20	15	0	1

If a primer row is double-clicked with the mouse or the "show alignment" button is pressed, the alignment of this primer against this genome is shown. The table and the alignment views can also be exported as text files.

BLAST result file: Bacillus\_cereus\_E33L vs test\_primer\_2

BLASTN 2.2.25+

Reference: Zheng Zhang, Scott Schwartz, Lukas Wagner, and Webb Miller (2000), "A greedy algorithm for aligning DNA sequences", J Comput Biol 2000; 7(1-2):203-14.

Database: G3.fna  
6 sequences; 5,843,235 total letters

Query= test\_primer\_2  
Length=20

Sequences producing significant alignments:	Score (Bits)	E Value
NC_007105	38.1	8e-005

> NC\_007105  
Length=53501

Score = 38.1 bits (20), Expect = 8e-005  
Identities = 20/20 (100%), Gaps = 0/20 (0%)  
Strand=Plus/Plus

```

Query 1      TATGTATCACCTGTATTAGA  20
           |||
Sbjct 19861  TATGTATCACCTGTATTAGA  19880

```

Lambda    K    H  
1.33    0.621    1.12

Gapped  
Lambda    K    H  
1.28    0.460    0.850

Effective search space used: 23372556

Database: G3.fna  
Posted date: May 24, 2012 11:02 AM  
Number of letters in database: 5,843,235  
Number of sequences in database: 6

Matrix: blastn matrix 1 -2  
Gap Penalties: Existence: 0, Extension: 2.5

Export
OK